Super Natural II—a database of natural products

Priyanka Banerjee[1,2], Jevgeni Erehman[1], Björn-Oliver Gohlke[1,3], Thomas Wilhelm[4], Robert Preissner[1,3] and Mathias Dunkel[1,*]

[1]Structural Bioinformatics Group, Charite-University Medicine Berlin, Institute of Physiology, Lindenberger Weg 80, 13125 Berlin, Germany, [2]Graduate School of Computational Systems Biology, Humboldt-Universität zu Berlin Invalidenstrasse 42, 10115 Berlin, Germany, [3]German Cancer Consortium (DKTK), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany and [4]Institute of Food Research, Norwich Research Park, Colney, Norwich, UK

ABSTRACT

Natural products play a significant role in drug discovery and development. Many topological pharmacophore patterns are common between natural products and commercial drugs. A better understanding of the specific physicochemical and structural features of natural products is important for corresponding drug development. Several encyclopedias of natural compounds have been composed, but the information remains scattered or not freely available. The first version of the Supernatural database containing ~50 000 compounds was published in 2006 to face these challenges. Here we present a new, updated and expanded version of natural product database, Super Natural II (http://bioinformatics.charite.de/supernatural), comprising ~326 000 molecules. It provides all corresponding 2D structures, the most important structural and physicochemical properties, the predicted toxicity class for ~170 000 compounds and the vendor information for the vast majority of compounds. The new version allows a template-based search for similar compounds as well as a search for compound names, vendors, specific physical properties or any substructures. Super Natural II also provides information about the pathways associated with synthesis and degradation of the natural products, as well as their mechanism of action with respect to structurally similar drugs and their target proteins.

INTRODUCTION

Natural products (NPs) are classical starting points for drug discovery. All herbal-based medicines are derived from natural compounds (1). NPs were involved in the development of ~64% of all drugs (2). For instance, easily available drugs such as lovastatin, paclitaxel, penicillin and sili-

binin were either directly or indirectly derived from NPs. NPs are an invaluable source of inspiration for organic chemists to synthesize novel drug candidates (3–6). Macrocycles, ring structures with more than 12 atoms, are a typical structural feature of natural compounds. They help to organize the overall molecule structure such that key functional groups can specifically interact with binding sites of target molecules such as proteins, resulting in few or no entropic loss on binding. Often macrocyclic NPs such as erythromycin, rapamycin, tacrolimus possess favorable drug-like physicochemical and pharmacokinetic properties like lipophilicity, metabolic stability, increased solubility and bioavailability (7). Such properties are important for protein–protein interaction targets (8). It is known that more NP-like molecules are needed for corresponding library design of the pharmaceutical industry (9,10). An interesting example for NP-based drugs starts with camptothecin isolated from the bark and stem of *Camptotheca acuminata* (11), which inhibits the DNA enzyme topoisomerase I. Because of low solubility and high adverse drug reaction during clinical trials (12), it was modified to the analogs topotecan and irinitecan (13). These analogs have been approved for cancer chemotherapy. The availability of NP databases is important for *in silico* screening in drug discovery. However, NP databases, such as the Dictionary of Natural Products (http://dnp.chemnetbase.com) and Natural Product Alert (14), are often commercial or freely available only with restricted information.

Better knowledge of the anabolism and catabolism of NPs leads to a relevant understanding of their ecological role. Often, primary metabolic pathways are single product and target-oriented (like tryptophan is synthesized in the tryptophan pathway, cholesterol is synthesized in the cholesterol pathway), whereas the pathways of the secondary metabolism are diversity-oriented, resulting in more than one product (15). This unique chemical diversity of secondary metabolites is one of the reasons for the continuing scientific interest in NPs (16,17,18). Nature has designed and selected secondary metabolites for different purposes. They often act as 'defense compounds' and were optimized

*To whom correspondence should be addressed. Tel: +49 30 450 540 755; Fax: +49 30 450 540 955; Email: mathias.dunkel@charite.de

during evolution for specific interactions with biological receptors (15,16).

Some NPs are toxic and produce adverse effects on cells or the whole organism. For instance, consumption of alpha-amanitin, a toxic peptide produced by *Amanita* mushrooms (19), can lead to irreversible kidney and liver damage (20). This NP molecule is considered as one of the deadliest compounds known.

Super Natural II is a freely available, web-based and easy to access database of NPs. Super Natural II contains a lot of information for each compound, including NP class, predicted toxicity class, mechanism of action (MoA) and pathways information. Super Natural II provides possible toxicity alert for the use of a particular natural compound. However, the absence of such toxicity prediction or alert for a compound should not be taken as an indication of safety. The toxicity prediction for the database compounds is based on their structural similarity to the known toxic compounds, calculated using ProTox (21). ProTox shows good performance in comparison to other commercial and free toxicity prediction methods.

To the best of our knowledge, Super Natural II is the first publicly available database of natural compounds with as many as 326 000 molecules and the corresponding wealth of additional information. Super Natural II provides information on MoA, pathways information and toxicity information integrated into one single platform. A comparison of Super Natural II to other NP databases is provided in Supplementary Table S1.

MATERIALS AND METHODS

The Super Natural II database contains ∼326 000 unique compounds, collected from 16 suppliers (as listed on the Super Natural II website) and five freely available databases (22–26).

Data preparation

The collected raw data from 21 sources were standardized using the chemoinformatic pipeline of the Konstanz Information Miner (KNIME) (27) (Steps 2–5) and ChemAxon (http://www.chemaxon.com). (Step 1): (1) standardizing files into the structure-data (SD) format, using JChem. In this process, incomplete structures were deleted, charges normalized and ions and smaller parts (e.g. water, salts) removed. (2) Reading standardized SD files, (3) generating InChIKeys (RDKit to InChi), (4) removing duplicates and (5) writing new SD files (see Supplementary Figure S3). Corresponding supplier information is mapped to each unique structure. The compounds are listed in the final database with a unique identifier (SN_ID). The 2D structure of each compound provided by the suppliers was used to generate 3D structures (Discovery studio, Accelrys Inc., http://www.accelrys.com/dstudio). A set of physicochemical descriptors such as molecular weight, log*P*, H-bond donors, H-bond acceptors, number of rings, aromatic rings, number of bonds and the number of heavy atoms was computed for each compound using JChem 6.1.3 (November 2013), ChemAxon (http://www.chemaxon.com).

Similarity search and substructure search

The structural similarity search was implemented using Open Babel (28). The Tanimoto coefficient (29) is used to measure the 2D similarity. It compares the structural similarity between the query molecule and the database entries using a concatenated fingerprint of the 'FP2' and 'FP4' Open Babel fingerprints calculated by Mychem (http://mychem.sourceforge.net). The Tanimoto coefficient is a number between 0 and 1 (1 corresponding to 'maximum similarity') (30). Pre-calculated fingerprints for all database entries are stored as blob objects in the MySQL-database. For a query structure it is calculated during the search. The top 15 results are returned.

For the substructure search, a fast-search index for all database compounds was created using Open Babel (28). The computed index is then queried by Simplified Molecular Input Line Entry Specification (SMILES) (31); resulting in up to 4000 hits per substructure search.

Classification

Due to the greater structural diversity of NPs compared to synthetic compounds, an increased therapeutic spectrum can be covered (32). Different classification schemes for NPs exist; Super Natural II has implemented a classification by structural characteristics such as alkaloid, amino acid or fatty acid.

Toxicity prediction

The toxicity class for each compound in the database was calculated using ProTox (21). The toxicity is categorized into six classes (ranging from classes I to VI) based on the lethal dose (LD50) values in mg/kg body weight in rodents. The toxicity class is predicted for ∼170 000 compounds; for unpredicted compounds (out of prediction range) the class is displayed as zero. For a better understanding of the toxicity prediction a link to the ProTox web server is provided. The number of NPs predicted for each toxicity class is reported in Supplementary Table S2.

Mechanism of action

Drug–target relations were taken from SuperTarget (33), which comprises >195 000 pharmacologically active compounds for which interactions are quantitatively known. These drugs were compared to all the Super Natural II compounds using a Tanimoto coefficient of 0.8 and above; the corresponding information was stored. Each NP was then mapped to its predicted target using the information from the drug-NP similarity mapping. It is generally accepted that structurally similar compounds have usually similar biological properties (34). The target prediction was computed using SuperPred web server (35), which has a prediction accuracy of 75.1%. Not all the NPs are yet assigned to a target, so similarity of a query NP to a ligand with known receptor association(s) may provide useful insight. For instance, search with target 'Bcl-2' and 'Apoptosis regulator Bcl-2' results in a list of drugs (15 hits) that are acting on this specific target. This drug could be a natural compound like 'Melatonin' (36); additionally, with the help of

pathway information on the results page, more specific facts on the pathways involved (such as Apoptosis, Alzheimer's disease) and corresponding species information can also be obtained.

Pathways

Considering information from KEGG (22) pathways often helps to better understand a specific NP–target protein interaction. In order to display pathway maps, NPs were mapped to potential targets with respect to their similarity to the reference drugs (see 'mechanism of action' section).

For instance, given a search for the organism 'Homo sapiens' and the pathway 'Apoptosis', the result page will display the corresponding pathway map with all known targets highlighted. A cursor-over on the target will display drugs acting on the target (up to 15 hits), together with similar NPs and related similarity score. Currently, the database is featuring all Homo sapiens pathways, as well as pathways from bacteria and fungi.

Clustering

A compound clustering based on structural similarity using Tanimoto coefficients was performed for the entire data set. The clustering was generated using an algorithm based on DBSCAN (density-based spatial clustering of application with noise) (37). The structural similarity clustering was calculated using similarity measures based on the concatenated Open Babel FP24 fingerprint, calculated using Mychem (http://mychem.sourceforge.net). The clustering was based on the corresponding similarity matrix with Tanimoto coefficient entries of 0.8 and above. The clustering is an effective tool to compare structural similarities between a large set of natural compounds. It results in ∼27 000 clusters with a minimum size of four compounds. Larger clusters are broken down into smaller subsets, containing at most 30 compounds in order to simplify the interpretation of the complex clustering results. The clustering results are visualized in an interactive heat map where red color corresponds to high similarity and light colors to low. A cursor-over on the heat map helps to view the corresponding compound structures.

Server, database and system requirements

Super Natural II is based on a relational MySQL database (http://www.mysql.com). To handle the chemical information within the database, the MyChem package (http://mychem.sourceforge.net/) is used. For most of its functions, MyChem relies on the Open Babel toolbox (http://openbabel.org). The Super Natural II website is built using PHP (http://www.php.net/), JavaScript (http://www.java.com/), Ajax; web access is enabled via an Apache HTTP Server (http://http.apache.org/). The site is best viewed in Mozilla Firefox and also functional on alternate browsers like Google Chrome and Microsoft Internet Explorer.

SEARCH OPTIONS

Natural compounds can be found in the Super Natural II database via several search methods. A web-based search

tool is available, incorporating a molecular drawing interface. Users can also search for compounds by properties, chemical structures or by a combination of criteria. Additionally, the integrated PubChem (38) search for a compound name is implemented. The search results page shows the main properties and the chemical structure of the compound, as well as a link to download the structure file in the Mol2 format. The web system was designed to allow a rapid and detailed response to specific queries. Important search options are as follows (Figure 1):

(i) First, a compound name or known supplier code can be used to retrieve information from the database. To get a subset of molecules having particular physicochemical properties, a search can be refined with more options like molecular weight, log$P$ value, hydrogen-bond acceptor, hydrogen-bond donor and classification. Additionally, one can also look for toxic natural compounds using the toxicity class filter.

(ii) Using the ChemDoodle (http://www.chemdoodle.com) sketcher, a user can build or import a molecular structure and perform a similarity or substructure search with compounds of the Super Natural II database.

(iii) Substructure search can also be done by pre-defined scaffolds via template search. Here, users can look for compounds in the database containing specific groups like certain amino acids, alpha sugars, beta sugars, D-sugars, aromatic rings, bicyclical rings, fused rings, heterocyclic or polycyclic rings, etc.

(iv) An MoA search enables users to screen NPs for information on their putative molecular targets and corresponding pathways. Similarly, the user can specify a target protein as search term (Uniprot ID/protein name) and get information on natural compounds which are predicted to interact with the specified target protein (see 'mechanism of action' section).

(v) The pathway search enables users to display known NPs and interacting protein targets associated with any pathway map from KEGG. The pathways can be searched by species, pathway type and function. Each protein target interacting with an NP is highlighted in yellow. A cursor-over option shows detailed information on similar compounds.

CONCLUSION

The Super Natural II database is a freely available resource with different embedded search functions. It allows a simple access to the growing number of available natural compounds. The compounds of the database comprise a rich chemical diversity and a wide spectrum of biological and pharmacological activities. Associated pathways and known and predicted mechanisms of action supply additional information. We expect that Super Natural II will be useful for studies involving virtual screening, metabolomics and the design of new compounds. The information provided by Super Natural II about structures and physicochemical properties of ∼326 000 NPs will contribute to future drug development.
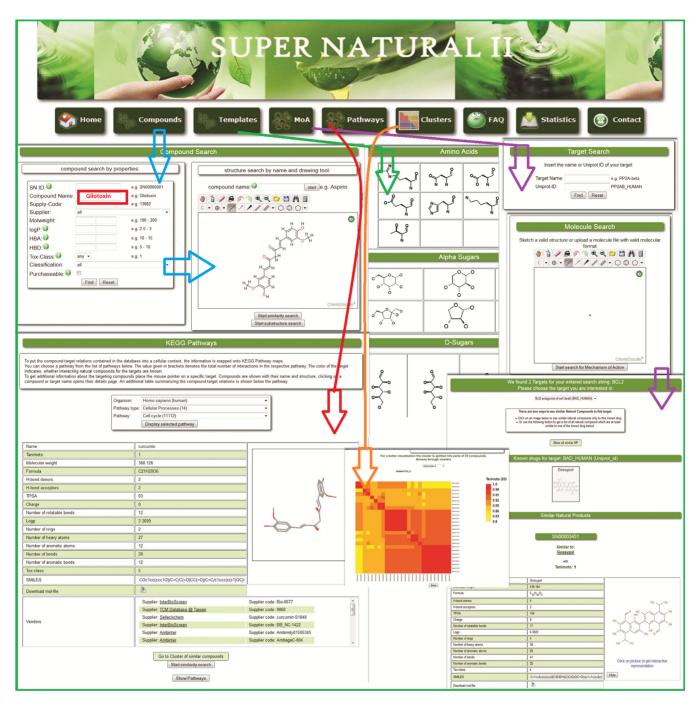
Figure 1. Super Natural II database search options. The compound search (blue arrow) enable search via compound name, supplier, molecular weight, log*P*, hydrogen-bond donor and acceptor, classifications, toxicity class as well as compounds that are purchasable. Template search (green arrow) allows searching via a pre-defined template. MoA search (purple arrow) can be performed for a molecule structure or a target name; clusters (orange arrow) can be searched via name, SNID (unique identifier for the Super Natural II database), IUPAC name or SMILES. All search results contain information about physicochemical properties, toxicity class, vendor and pathway. Cluster results are displayed using heat maps and related similarity values for compounds of the cluster. Pathway information (red arrow) can be obtained for specific species and pathway. Additionally, the molecule structure is shown.

AVAILABILITY

Super Natural II is available through the website without registration: http://bioinformatics.charite.de/supernatural.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

REFERENCES

1. Bent,S. (2008) Herbal medicine in the United States: review of efficacy, safety, and regulation: grand rounds at University of California, San Francisco Medical Center. *J. Gen. Intern. Med.*, **23**, 854–859.
2. Newman,D.J. and Cragg,G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.*, **75**, 311–335.
3. Beghyn,T., Deprez-Poulain,R., Willand,N., Folleas,B. and Deprez,B. (2008) Natural compounds: leads or ideas? Bioinspired molecules for drug discovery. *Chem. Biol. Drug Des.*, **72**, 3–15.
4. Koehn,F.E. and Carter,G.T. (2005) The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.*, **4**, 206–220.
5. Cragg,G.M. and Newman,D.J. (2013) Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta*, **1830**, 3670–3695.
6. Costantino,L. and Barlocco,D. (2006) Privileged structures as leads in medicinal chemistry. *Curr. Med. Chem.*, **13**, 65–85.
7. Driggers,E.M., Hale,S.P., Lee,J. and Terrett,N.K. (2008) The exploration of macrocycles for drug discovery-an underexploited structural class. *Nat. Rev. Drug Discov.*, **7**, 608–624.
8. Cross,S., Baroni,M., Ortuso,F., Alcaro,S. and Cruciani,G. (2013) Disrupting protein-protein interfaces using GRID molecular interaction fields. In: Mangani,S (ed). *Disruption of Protein-Protein Interfaces SE - 3*. Springer, Berlin Heidelberg, Germany, pp. 61–82.
9. Aubé,J. (2012) Small-molecule libraries: naturally inspired oligomers. *Nat. Chem.*, **4**, 71–72.
10. Vazhayil,B.K., Sundaram,R.S., Annapandian,V.M., Abhirama,B.R., Sudha,M. and Thiyagarajan,T. (2014) Natural products and its derived drugs for the treatment of neurodegenerative disorders: Alzheimer's Disease-A. *Br. Biomed. Bull.*, **2**, 359–370.
11. Efferth,T., Fu,Y.-J., Zu,Y.-G., Schwarz,G., Konkimalla,V.S.B. and Wink,M. (2007) Molecular target-guided tumor therapy with natural products derived from traditional Chinese medicine. *Curr. Med. Chem.*, **14**, 2024–2032.
12. Wall,M.E., Wani,M.C., Cook,C.E., Palmer,K.H., McPhail,A.T. and Sim,G.A. (1966) Plant antitumor agents. I. The isolation and structure of camptothecin, a novel alkaloidal leukemia and tumor inhibitor from Camptotheca acuminata. *J. Am. Chem. Soc.*, **88**, 3888–3890.
13. Ulukan,H. and Swaan,P.W. (2002) Camptothecins: a review of their chemotherapeutic potential. *Drugs2002*, **62**, 2039–2057.
14. Loub,W.D., Farnsworth,N.R., Soejarto,D.D. and Quinn,M.L. (1985) NAPRALERT: computer handling of natural product research data. *J. Chem. Inf. Comput. Sci.*, **25**, 99–103.
15. Fischbach,M.A. and Clardy,J. (2007) Commentary: one pathway, many products. *Nat. Chem. Biol.*, **3**, 353–355.
16. Li,J.W.-H. and Vederas,J.C. (2009) Drug discovery and natural products: end of an era or an endless frontier? *Science*, **325**, 161–165.
17. Newman,D.J. and Cragg,G.M. (2009) Natural product scaffolds as leads to drugs. *Future Med. Chem.*, **1**, 1415–1427.
18. Grabowski,K. and Schneider,G. (2007) Properties and architecture of drugs and natural products revisited. *Curr. Chem. Biol.*, **1**, 115–127.
19. Hallen,H.E., Luo,H., Scott-Craig,J.S. and Walton,J.D. (2007) Gene family encoding the major toxins of lethal Amanita mushrooms. *Proc. Natl Acad. Sci. U.S.A.*, **104**, 19097–19101.
20. Yilmaz,I., Kaya,E., Sinirlioglu,Z.A., Bayram,R., Surmen,M.G. and Colakoglu,S. (2014) Clinical importance of toxin concentration in Amanita verna mushroom. *Toxicon*, **87**, 68–75.
21. Drwal,M.N., Banerjee,P., Dunkel,M., Wettig,M.R. and Preissner,R. (2014) ProTox: a web server for the in silico prediction of rodent oral toxicity. *Nucleic Acids Res.*, **42**, W53–W58.
22. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
23. Caspi,R., Altman,T., Billington,R., Dreher,K., Foerster,H., Fulcher,C.a., Holland,T.a., Keseler,I.M., Kothari,A., Kubo,A. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459–D471.
24. Gu,J., Gui,Y., Chen,L., Yuan,G., Lu,H.-Z. and Xu,X. (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE*, **8**, e62839.
25. Wishart,D.S., Jewison,T., Guo,A.C., Wilson,M., Knox,C., Liu,Y., Djoumbou,Y., Mandal,R., Aziat,F., Dong,E. *et al.* (2013) HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801–D807.
26. Irwin,J.J., Sterling,T., Mysinger,M.M., Bolstad,E.S. and Coleman,R.G. (2012) ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, **52**, 1757–1768.
27. Berthold,M., Cebron,N., Dill,F., Gabriel,T., Kötter,T., Meinl,T., Ohl,P., Sieb,C., Thiel,K. and Wiswedel,B. (2008) KNIME: The Konstanz Information Miner. In: Preisach,C, Burkhardt,H, Schmidt-Thieme,L and Decker,R (eds). *Data Analysis, Machine Learning and Applications SE - 38*. Springer, Berlin Heidelberg, Germany, pp. 319–326.
28. O'Boyle,N.M., Banck,M., James,C.a., Morley,C., Vandermeersch,T. and Hutchison,G.R. (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33–47.
29. Delaney,J.S. (1996) Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Divers.*, **1**, 217–222.
30. Martin,Y.C., Kofron,J.L. and Traphagen,L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.
31. Weininger,D. (1988) SMILES a chemical language and information system. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
32. Dunkel,M., Fullbeck,M., Neumann,S. and Preissner,R. (2006) SuperNatural: a searchable database of available natural compounds. *Nucleic Acids Res.*, **34**, D678–D683.
33. Hecker,N., Ahmed,J., von Eichborn,J., Dunkel,M., Macha,K., Eckert,A., Gilson,M.K., Bourne,P.E. and Preissner,R. (2012) SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.*, **40**, D1113–D1117.
34. Keiser,M.J., Irwin,J.J. and Shoichet,B.K. (2010) The chemical basis of pharmacology. *Biochemistry*, **49**, 10267–10276.
35. Nickel,J., Gohlke,B.-O., Erehman,J., Banerjee,P., Rong,W.W., Goede,A., Dunkel,M. and Preissner,R. (2014) SuperPred: update on drug classification and target prediction. *Nucleic Acids Res.*, **42**, W26–W31.
36. Luchetti,F., Canonico,B., Betti,M., Arcangeletti,M., Pilolli,F., Piroddi,M., Canesi,L., Papa,S. and Galli,F. (2010) Melatonin signaling and cell protection function. *FASEB J.*, **24**, 3603–3624.
37. Yue,S., Li,P., Guo,J. and Zhou,S. (2004) Using Greedy algorithm: DBSCAN revisited II. *J. Zhejiang Univ. Sci.*, **5**, 1405–1412.
38. Wang,Y., Suzek,T., Zhang,J., Wang,J., He,S., Cheng,T., Shoemaker,B.a., Gindulyte,A. and Bryant,S.H. (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, **42**, D1075–D1082.